DOCUMENT RESUME

ED 232 219 CS 504 246

AUTHOR Stano, Michael
TITLE The Critical Incident Technique: A Description of the Method.
PUB DATE Apr 83
NOTE 15p.; Paper presented at the Annual Meeting of the Southern Speech Communication Association (Lincoln, NE, April 7-9, 1983).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Behavior Patterns; Communication Research; *Critical Incidents Method; Data Analysis; Evaluation Methods; *Research Design; *Research Methodology; Speech Communication

ABSTRACT

Critical Incident Technique (CIT) involves the collection of real-world examples of behavior that characterize either very effective or very ineffective performance of some activity. The principal advantage of the CIT is that it generates data based on actual behavior rather than on a particular researcher's subjectivity. The CIT has much to offer speech communication researchers, especially those in the applied communication area where the stress is on what actually works in real-world settings. Using the CIT, the research should collect and interpret examples of behavior from a sample before reviewing much of the existing literature on the subject. Questionnaire respondents should be told to think of recent examples related to the phenomenon being studied that stand out in their minds. These incidents are then subject to the equivalent of a content or factor analysis, sorted into a comprehensive list of behaviors. The researcher may elect to end the study and report or act on the findings, or he or she may take an additional step. To ensure that all distinct behaviors in the incidents were identified, to ensure there are no overlapping or repetitive categories, and to check on the appropriateness of the summary statements, the categories must be taken to either the original or a new set of subjects. If a more precise follow up is desired, the researcher may then examine methods developed by other researchers. (HTH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED232219

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

THE CRITICAL INCIDENT TECHNIQUE: A DESCRIPTION OF THE METHOD

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Michael Stano
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Michael Stano

109 Morrill Hall
Oklahoma State University
Stillwater, OK 74078
phone: 405-624-6150

Michael Stano (Ph.D., University of Minnesota, 1977) is an associate professor of Speech Communication at Oklahoma State University, Stillwater, OK 74078.

Running Head: CIT

CS 504 246

ABSTRACT

The Critical Incident Technique (CIT) is a method of research which holds much promise for speech communication scholars, especially those interested in applied communication strategies. The purpose of this paper is to describe how the CIT works so that those not familiar with it can begin to use it in research.

THE CRITICAL INCIDENT TECHNIQUE: A DESCRIPTION OF THE METHOD

Introduction

Almost 40 years ago, John C. Flanagan began developing and writing about the Critical Incident Technique (CIT). In essence, the CIT involves the collection of real-world examples of behavior which characterize either very effective or very ineffective performance of some activity.

For the most part, critical incident methodology has been employed by personnel specialists to outline important job duties and skills (Fivars, 1975). In an early review of applications, Flanagan (1954, p. 346) indicated that the CIT had been used to describe typical performance, to develop evaluation criteria, to formulate training programs, to help in personnel selection and classification, to assist in job design and purification, to aid in equipment design, to determine operating procedures, to assess levels of motivation and attitudes towards leadership, and in counseling and psychotherapy. While dated, these categories of applications encompass nearly all of the CIT research which has been done to date.

The principle advantage of the CIT is that it generates data which is based on actual behavior rather than on a particular researcher's inferences, hunches, stereotypes, and subjective estimates of what is important or necessary (Flanagan, 1949a, pp. 39-40, 48, 53; 1949b, pp. 39, 42; 1951, p. 29; 1954, pp. 329, 355). To avoid the subjective is especially critical when developing selection procedures and performance appraisal schemes; the Equal Employment Opportunity Commission and other agencies which work against discriminatory employment practices take a dim view of any selection or evaluation method which is not job based and behaviorally

grounded. The CIT is useful in that it is specifically designed to minimize general impressions of irrelevant personal factors and maximize systematic observations.

So far, very few communication scholars have taken advantage of the CIT. The CIT has much to offer speech communication researchers, especially those in the applied communication area where the stress is on what actually works in real-world settings. The CIT could be employed to provide paradigms of effective communication practice in many different arenas and, seen through the lens of the CIT, behaviors required for successful communication could take on new sharpness. The purpose of this paper is to describe precisely how the CIT works so that those not familiar with it can begin to use it in research.

A Description of the CIT

Flanagan (1954) notes:

> It should be emphasized that the Critical Incident Technique does not consist of a single rigid set of rules governing . . . data collection. Rather it should be thought of as a flexible set of principles which must be modified and adapted to meet the specific situation at hand. (p. 335)

This notion should be kept in mind as the following procedures are reviewed.

Flanagan recommends that the gathering of critical incidents begin before the researcher has reviewed much of the literature pertaining to the topic to be studied. This practice is somewhat unorthodox but it does make sense. Using the CIT, the researcher must collect and interpret examples of behavior from some sample. This collection and interpretation process may become biased if too much is known about the behavior in advance. That is, the investigator may be led to see in the behaviors which are submitted only that which previous investigators have found

important. As Flanagan (1954) explains it:

> An effective strategy is, at first, to literally ignore the literature of theory and fact on the area under study, in order to assure that the emergence of categories will not be contaminated by concepts more suited to different areas. Similarities and convergences can be established after the analytic categories have emerged. (p. 337)

The first step in building a critical incident questionnaire is the selection of a frame of reference for the respondent. This preliminary requirement is clarified by Flanagan (1954):

> A basic condition necessary for any work on the formulation of a functional description of an activity is a fundamental orientation in terms of the general aims of the activity. . . . The most useful statements of aims seem to center around some simple phrase or catchword which is slogan-like in character. Such words provide a maximum of communcation with only a minimum of possible misinterpretation. Such words as "appreciation," "efficiency," "development," "production," and "service" are likely to be prominent in statements of general aims. (pp. 336-337)

The catchword and the statement of general aims of the activity will be used in the instructions subjects receive. While the author of this paper knows of no research in the area, it is very possible that different catchwords might elicit different behavioral examples from respondents. For example, respondents asked to provide critical incidents of "effective listening" might suggest different items if asked to provide critical incidents of "productive listening." As the frame of reference given the subject changes, so too might the data which is produced. Consequently, researchers should carefully consider the rhetoric selected to describe the activity.

Based on the catchword, instructions to subjects must be formulated. In general, respondents should be told to think of recent examples related to the phenomenon being studied which stand out in their mind. Subjects might, for example, be asked to recall their last performance appraisal interview and to describe several specific things their supervisor did to

make the interview productive or unproductive. The idea is to collect examples of behavior in the form of anecdotes or stories that pertain to the most important aspects of the activity; these reported examples should not concern routine behaviors but rather behaviors which make for success or failure in the activity (Flanagan, 1949a, pp. 40, 45-56; Kirchner and Dunnette, 1971, p. 359). Usually, subjects are asked to provide incidents related to both success and failure.

Incidents may be collected either in interviews or via questionnaires. Flanagan seems to prefer the interview approach (1954, p. 340) but the time involved and the difficulties in recording orally recited incidents make the questionnaire more feasible; many subjects may be canvassed simultaneously and the incidents are individually written in the subject's own words. Andersson and Nilsson (1964) have found that subjects in interviews provide more incidents (5.0 versus 2.5) and that many are prone to not return questionnaires; despite this, there were no real differences in incidents reported by interview and questionnaire groups. Regardless of which data collection method is used, respondents must be convinced of total anonymity or their honesty and the validity of the incidents will be suspect (Flanagan, 1954, p. 341).

A most important step in the CIT is the selection of a sample. Those who are asked to provide critical incidents should definitely be in a position to observe the behavior in question. Since the CIT is behaviorally grounded, subjects must have the opportunity to see others engaged in the activity being studied. In studies of work performance, the supervisors of the job under investigation, or others who are job knowledgeable, would be appropriate observers (Flanagan, 1949a, pp. 43,45).

Once the collection of critical incidents begins, so does one of the

researcher's most onerous tasks: the sorting of incidents into common-theme categories (Flanagan, 1949a, pp. 47-48; 1954, pp. 344-345). The objective is to digest from the many incidents a comprehensive list of the behaviors mentioned. The process is analogous to a content or factor analysis.

Initially, the researcher must separate each unique incident mentioned by each subject. Whether working from an interview transcript or recording or from a questionnaire, the researcher should record on notecards every identifiable behavior. Some subjects may mention the same behavior as critical; the sorting project begins by grouping cards containing obviously identical incidents together. The researcher then moves to cards where the sorting decision is less clear-cut. Because of idiosyncratic language useage, lack of clarity in expression, semantics, and other ambiguities, it may not be clear if two incidents are really descriptions of the same behavior; if they are, they should be put in the same category and if they are not, they belong in distinct categories. This phase of the sorting sounds simpler than it is; this is a crucial step in the CIT and it is one which, according to Flanagan (1954, p. 344), requires "insight, experience, and judgment." Throughout this step, cards mentioning successes and failures are put in the same stack if they seem to concern the same behavior.

In the end, the researcher will have a number of piles of cards representing categories and subcategories. Single sentence summary statements must be composed for each of these piles (Flanagan, 1954, p. 345). These summaries attempt to cut through variations in terminology in order to capture the essence of the activity being discussed. The heading should be neutral rather than defining unsatisfactory or outstanding behaviors and

they should be phrased in positive terms.

At this juncture it is possible to discuss the question of how many incidents must be collected. Flanagan is less than specific on this point (Flanagan, 1954, p. 343). He indicates that anywhere from 50 to 4,000 incidents may be needed to provide a complete description of job requirements. To get an indication of when it is permissible to stop collecting data, Flanagan suggests the researcher do a tentative sort of the incidents which have been gathered. The researcher should then look through a new set of 100 cards and see if the category system has an appropriate spot for each. If the new batch of cards contains only two or three previously unmentioned behaviors, the researcher may consider the coverage adequate. If there are more than three unplaceable items, the categories must be revised and tested by another 100 cards. Certainly, many groups with which investigators work are not capable of generating even 100 incidents, let alone several hundred or several thousand; individual investigators must use their own judgment on how many incidents are realistic and they must, on a case-to-case basis, consider any limitations on validity stemming from sample size.

When defects in the category system have been repaired, the researcher may elect to end the study and report or act on the findings. Alternately, an additional step may be undertaken. This step is one which acknowledges the subjectivity and fallibility of the sorting procedure and seeks corrections. To make sure that all distinct behaviors in the incidents were identified, to make sure there are no overlapping or repetitive categories, and to check on the appropriateness of the summary statements, the cards and categories may be taken to either the generators of the incidents or to a new group of subjects. Flanagan discusses this check-back procedure in

very simple qualitative terms: one goes to the reviewers and asks whether in their experience the list of behaviors are those critical for the activity (Flanagan, 1951, p. 31). The investigator will give consideration to the feedback received in compiling a final report.

If more precision in the follow-up is desired, the researcher must leave the works of Flanagan and examine methods developed by others. Andersson and Nilsson (1964), for example, selected at random 100 incidents from their original set of 1,847; these items were given to 24 psychology students along with the category labels which had been developed. Working in pairs, the students tried to sort the 100 incidents into their corresponding categories. A percentage of agreement with the original categorization was calculated. If the level of agreement is high, the researcher may place some confidence in the list of behaviors which evolved from the initial card sorting. The logic of this assumption is fairly straightforward (though unexplained by Andersson and Nilsson). If explicit category themes have been developed and if all identifiable behaviors have been correctly separated, the sorting task should be easy and there should be a great deal of correspondence between the original and new classification of cards. When the level of disagreement reaches a high level, a completeness or a redundancy or a phrasing fault is indicated. Discrepancies would surface only when sorters had been consciously or unconsciously forced to engage in random or strained assignment of items due to the imprecise nature of the paradigm being employed.

Personnel psychologists working on behaviorally anchored rating scales (BARS) have delineated other more quantitative techniques which nicely compliment the procedures outlined by Flanagan (Schwab, Heneman, and DeCotiis, 1975). The development of BARS involves collecting critical incidents and sorting them into categories in the usual way. To analyze the item-category

fit, another group of job knowledgeable participants tries to sort the incidents into the categories. This step is labeled "Retranslation" and rather than resulting in a revamping of the category system, it results in throwing out critical incidents that are not sorted reliably. Typically, a 50 to 80% agreement in the way the judges assign items to dimensions is expected or the item is not retained. Following retranslation, the second group of participants is requested to rate, on a seven- or nine-point scale, the degree to which each incident represents effective or ineffective performance on the appropriate dimension. The standard deviation of each item is then calculated; if the judges agree on the effectiveness level described in the incident, a low standard deviation will result. Any item which, on a seven-point scale, has a standard deviation of more than 1.50 will usually be thrown out. The final BARS instrument contains only those items that survived the retranslation and standard deviation tests. The items for each dimension are placed on scales and arranged in the order determined by the previous rating step. In evaluating performance, observers simply check the position on the scale which most closely corresponds to the behavior of the person being watched. The danger in the BARS method is that in throwing out an item with a low agreement level or a large standard deviation one might be removing from the data set a unique and critical behavior which was badly sorted due to poor category formation on the part of the researcher. While this method may purify the categories and make them homogeneous, it does not assure the validity or completeness of the total category system.

The last category-checking scheme to be considered was developed by Stano (1977). After collecting incidents and forming categories, a single volunteer was handed cards containing all of the critical incidents and a list of the categories. Once the individual finished sorting, the number

of times the sorter and the researcher disagreed on card placement was counted and a Scott's Pi reliability coefficient was calculated (see Holsti, 1969, p. 140). Each dissimilarity was then examined. The researcher and the sorter explained to one another why they had put the item where they did and this process resulted in the combination of several previously separate categories, the creation of new categories, and the rephrasing of a few general summary statements. At the end of the error-examination session, 100% agreement was present regarding the language of the category labels and the distribution of behaviors into the categories.

A second sorter was given not the researcher's original categories but rather the new and improved system worked out in conjunction with the first sorter. The same procedures outlined above were followed and labels and card assignments were again revised.

A third sorter received the third edition of the material and progressed through identical stages. As a final check on clarity and completeness, two additional volunteers were asked merely to sort the cards and no adjustments were made.

These procedures resulted in Scott's Pi figures which marched steadily upward from .74 at the start to .91 at the end. One plausible explanation for this increased reliability is that the category labels became more lucid and understandable and more unique behaviors had been legitimately separated.

Stano (1977) took an additional step to assure the validity of the category system. A new sample of participants was drawn from the same population used to generate critical incidents. Participants were given an opportunity to observe someone engaging in the activity being studied and they were then asked to evaluate that individual's overall performance on a six-point productive-unproductive continuum. Further, subjects were given each of

the critical incident category labels and instructed to rate, on a seven-point scale ranging from "inaccurate" to "accurate," the degree to which each statement described the behavior of the person they had just observed. As a final task, subjects rated, on a seven-point scale, the importance of the category statements vis-a-vis the activity.

The collection of the various estimates just mentioned made possible two independent checks on the validity of the category system. One check involved the judgment of each statement's importance; if the category system developed from the critical incidents was on target, all items should receive high importance scores. The second check used the productive-unproductive rating as the independent variable and the accurate-inaccurate rating as the dependent variable; a one way ANOVA with a LSD aposteriori test (see Kim and Kohout, 1975, pp. 398-433) was performed to determine which behaviors were viewed as productive and which were viewed as unproductive. If a particular category of critical incidents is truly critical, the ANOVA should reveal a strong link between the category and the productive-unproductive evaluation; if the behavior described in the statement is not critical, no linkage should be present.

What Stano (1977) found using these procedures is not relevant to the present discussion. It is only important that the process of quantitatively checking the worth of the categories be understood.

The foregoing list of variations on Flanagan's original methodology is not meant to be exhaustive. Rather, it is merely illustrative of both simple and complex procedures which the researcher might use to enhance the quality of the subjectively derived category set.

Conclusion

In the introduction, it was asserted that the CIT is a method of research that should prove valuable to the applied communication scholar. Perhaps now that the CIT has been described that claim can more easily be accepted and uses of the CIT can more easily be discerned. Too often, researchers jump into testing specific and limited hypotheses without fully understanding the essential elements of the behavior they are studying. In the CIT there is a remedy for this cart-before-the-horse problem. Regarding the work of researchers in his own field Flanagan (1949a) has observed:

> Customarily the personnel psychologist has spent ninety to ninety-five per cent of his time trying to test hypotheses without having adequately defined his problem. It is suggested that a reversal of these emphases is in order. If the personnel psychologist will spend ninety to ninety-five per cent of his time collecting primary data regarding job behavior, the solutions to his other problems may not require more than the five or ten per cent of his time which remains. (p. 40).

The same suggestion may be equally relevant to those studying communication phenomena.

REFERENCES

Andersson, B. E., and Nilsson, S. G. Studies on the Reliability and Validity of the Critical Incident Technique, Journal of Applied Psychology, 1964, 48, 398-403.

Fivars, G. The Critical Incident Technique: A Bibliography. JSAS Catalog of Selected Documents in Psychology, 1975, 5, 210.

Flanagan, J. C. Job Requirements. In Current Trends in Industrial Psychology. Pittsburgh: University of Pittsburgh Press, 1949. (a)

Flanagan, J. C. A New Approach to Evaluating Personnel. Personnel, 1949, 26, 35-42. (b)

Flanagan, J. C. Defining the Requirements of the Executive's Job. Personnel, 1951, 28, 28-35.

Flanagan, J. C. The Critical Incident Technique. Psychological Bulletin, 1954, 51, 327-358.

Holsti, O. R. Content Analysis for the Social Sciences and Humanities. Reading, Mass.: Addison-Wesley, 1969.

Kim, J., and Kohout, F. J. Analysis of Variance and Covariance: Subprograms ANOVA and Oneway. In N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent (Eds.), Statistics Package for the Social Sciences (2nd ed.). New York: McGraw-Hill, 1975.

Kirchner, W. K., and Dunnette, M. D. Identifying the Critical Factors in Successful Salesmanship. In G. A. Yukl and K. N. Wexley (Eds.), Readings in Organizational and Industrial Psychology. New York: Oxford University Press, 1971.

Schwab, D. P., Heneman, H. G., III, and DeCotiis, T. A. Behaviorally Anchored Rating Scales: A Review of the Literature. Personnel Psychology, 1975, 28, 549-562.

Stano, M. E. Dimensions of Productive or Unproductive Performance Appraisal Interviews. Unpublished Doctoral Dissertation, University of Minnesota, 1977.